Gaussian Processes for Survival Analysis

by Tamara Fernández, Nicolás Rivera and Yee Whye Teh

Michael Komodromos

February 26, 2021

Table of Contents

1. Recap
   - Survival Analysis
   - Gaussian Processes
2. Gaussian Processes for Survival Analysis
   - Overview
   - Model
   - Adding Covariates
   - Inference
   - Censoring and Approximations
   - Experiments

Recap: Survival Analysis

Survival Analysis models time to failure events. Let T be a random variable denoting a time to failure event with pdf $f(t)$, then:

Survivor Function: $S(t) = 1 - F(t)$

Hazard Rate: $\lambda(t) = f(t)/S(t)$

Recap: Survival Analysis Identities

Some useful identities

$$\lambda(t) = \frac{f(t)}{S(t)} \tag{1}$$

$$= -\frac{S'(t)}{S(t)} \tag{2}$$

$$= -\frac{d}{dt}\log\left(S(t)\right) \tag{3}$$

Re-arranging Eq. (3) gives

$$S(t) = \exp\left(-\int_0^t \lambda(s)ds\right) \tag{4}$$

Recap: Gaussian Processes

Gaussian Processes: a collection of random variables, any finite number of which have a joint Gaussian dist.

A GP is specified by it's mean function $m(x) = \mathbb{E}[f(x)]$ and kernel function $k(x, x') = \mathbb{E}[(f(x) - m(x))(f(x') - m(x')]$.

We write

$$f(x) \sim GP(m(x), k(x, x')) \tag{5}$$

or just $f \sim GP$

Recap: Gaussian Processes

Suppose

$$y_i = f(X_i) + \epsilon_i, \tag{6}$$

where $\epsilon_i \sim N(0, \sigma^2)$.

We are interested in describing the uncertainty of f. Under a Bayesian framework we have

$$p(f|X, y) \propto p(f|X)p(y|f, X). \tag{7}$$

Recap: Gaussian Processes

$$p(f|X, y) \propto p(f|X)p(y|f, X). \tag{8}$$

If we place a GP prior over f, then we have

$$f|X \sim N(m(x), K(X, X)) \tag{9}$$

and

$$y|f, X \sim N(m(x), k(X, X) + \sigma^2 I_n) \tag{10}$$

Given our likelihood and prior are normally dist we know by conjugacy that $f|X, y \sim N$

Onto the paper!

A one slide overview

Semi-parametric method where the hazard rate

$$\lambda(t) \quad = \quad \underbrace{\lambda_0(t)}_{\text{parametric}} \quad \times \quad \underbrace{\sigma(l(t))}_{\text{non-parametric}} \tag{11}$$

where $\sigma(\cdot)$ is a link-function.

In other words, we specify the baseline hazard rate $\lambda_0(t)$ from a parametric distribution (Weibull, Exp etc...) and we model $l(\cdot)$ via a GP.

Model

We're interested in modelling time to failure events $T \in \mathbb{R}^+$ which has density $f(t)$, survivor function $S(t)$ and hazard rate $\lambda(t) = f/S$.

Model (without covariates)

$$I(\cdot) \sim GP(0, k), \quad \lambda(t)|I, \lambda_0 = \lambda_0(t)\sigma(I(t)), \quad T_i|\lambda \overset{iid}{\sim} f(t) \tag{12}$$

where $\sigma(x) = (1 + e^{-x})^{-1}$

Note

$$f(t) = \lambda(t) \exp\left(-\int_0^t \lambda(s)ds\right) \tag{13}$$

Interpretation

Let's look into $\lambda(t)$ a bit more. Under our model definition we have

$$\lambda(t) = \lambda_0(t)\sigma(l(t)) \tag{14}$$

So $\lambda_0(t)$ is derived from a parametric distribution and $\sigma(l(t))$ adjusts the baseline hazard rate by some multiplicative term. Noting that $0 \leq \sigma(x) \leq 1$.

Interpretation

Example

If we believe the baseline hazard rate is given by $1/\mu$ i.e. is the hazard rate of $\text{Exp}(1/\mu)$ recalling

$$\frac{f(t)}{S(t)} = \frac{f(t)}{1 - F(t)} = \frac{1}{\mu} \frac{e^{-t/\mu}}{(1 - (1 - e^{-t/\mu}))} = \frac{1}{\mu}. \tag{15}$$

Then our model adjusts the baseline hazard by $\sigma(l(t))$ for some $t > 0$

Interpretation

Example

If we believe the baseline hazard rate is given by $\beta t^{\alpha-1}$ for $\alpha, \beta > 0$, the hazard rate of a Weibull dist, then we're adjusting the baseline hazard by $\sigma(l(t))$.

How much adjustment are we expecting?

Let's consider, $E\left[\sigma(X)\right]$ where $X \sim N(0,1)$.

Result: MacLaurin expansion of $\sigma(x)$

$$\sigma(x) = \frac{1}{2} + \frac{1}{4}x - \frac{1}{48}x^3 + \cdots \tag{16}$$

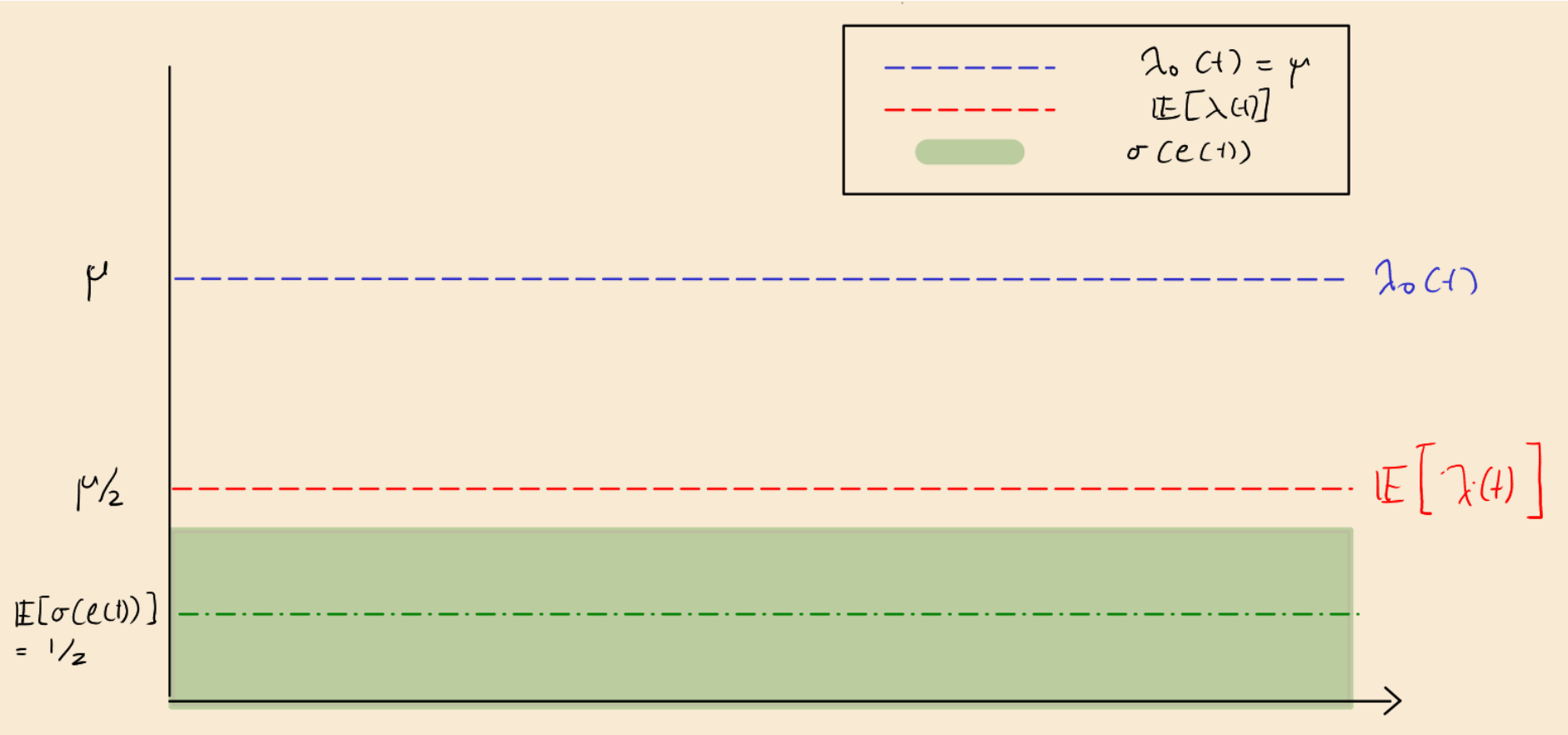
Now consider

$$\begin{aligned}
\mathbb{E}\left[\sigma(X)\right] &= \int_{\mathbb{R}} \sigma(x) f_X(x) dx && (17)\\
&= \int_{\mathbb{R}} \left(\frac{1}{2} + \frac{1}{4}x - \frac{1}{48}x^3 + \cdots\right) f_x(x) dx && (18)\\
&= \frac{1}{2} && (19)
\end{aligned}$$

Recall, $\mathbb{E}\left[(X-\mu)^p\right] = 0$ when p is odd.

Visualisation

Figure: Visualisation where we take the baseline hazard $\lambda_0(t) = \mu$



Making sure the model defines a well-behaved Survival function.

Proposition 1.

Let $(l(t))_{t\geq 0} \sim GP(0, k)$ be a stationary Gaussian Process. Suppose $k(s)$ is non-increasing and $\lim_{s\to\infty} = 0$. Further, assume there exists $K > 0$ and $\alpha > 0$ such that $\lambda_0(t) > Kt^{\alpha-1}$ for $t \geq 1$.

Then for random survival function $S(t)$ associated with $l(t)$ we have $\lim_{t\to\infty} S(t) \xrightarrow{p} 0$.

Where $\xrightarrow{p}$ denotes convergence in probability. A proof is given in the supplementary materials.

Any questions so far?

Adding Covariates

Covariates are introduced into the model through the kernel. Recall that we can construct kernels by performing basic operations (addition, multiplication) on multiple kernels.

Adding Covariates

Let $X \in \mathbb{R}^d$ be our covariates and t be the observed failure time, then for pairs (t, X) and (s, Y) we have

$$K((t, X), (s, Y)) = K_0(t, s) + \sum_{i=1}^{d} X_i Y_i K_i(t, s) \tag{20}$$

Model with Covariates

Assuming we have some covariates $X_i \in \mathbb{R}^d$, then the new model is given by

Model (with covariates)

$$l(\cdot) \sim GP(0, K), \quad \lambda_i(t)|l, \lambda_0(t), X_i = \lambda_0(t)\sigma(l(t, X_i))$$

$$T_i|\lambda \overset{\text{ind}}{\sim} \lambda(T_i)e^{-\int_0^{T_i} \lambda_i(s)ds}$$

We need to make sure of is that K is stationary

Inference

In general λ_i is not analytically tractable as λ_i is defined by a Gaussian Process.

Numerical methods can be used to approximate $\int \lambda(t)dt$ but these are computationally expensive.

So we step into a Poisson Process framework and use a data-augmentation scheme based on **Poisson thinning**.

Poisson Processes

Non-homogeneous Poisson Process

A non-homogeneous Poisson process is defined by it's intensity function, and has a probability mass function given by

$$P(N = n|\lambda(t)) = \frac{\Lambda(t)^n}{n!} e^{-\Lambda(t)} = \frac{S(t)\Lambda(t)^n}{n!} \tag{21}$$

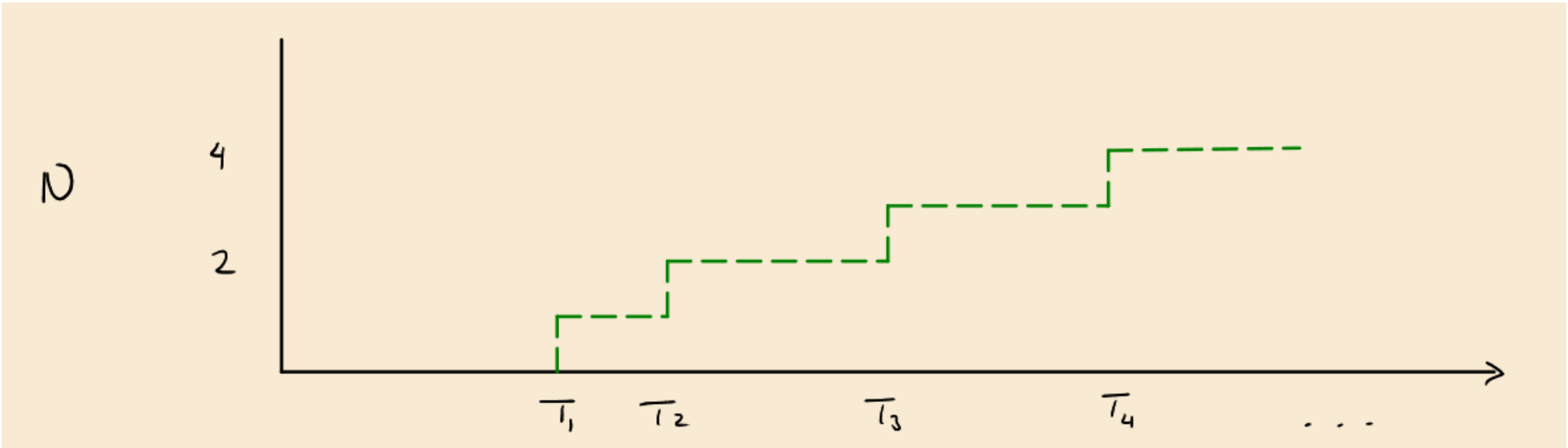
where $\Lambda(t) = \int_0^t \lambda(s)ds$

Differentiating Eq. (21) wrt. t we have

$$f_n(t) = \frac{f(t)\Lambda(t)^{n-1}}{(n-1)!} \tag{22}$$

which describes the density of the nth jump time.

Poisson Processes



Poisson Processes and Survival Analysis

Using

$$f_n(t) = \frac{f(t)\Lambda(t)^{n-1}}{(n-1)!} \tag{23}$$

and taking $n = 1$ we obtain the density of the first jump time T of a Non-homogeneous Poisson process is $f(t)$.

We recall,

$$p(T_i|\lambda) = \lambda(T_i)e^{-\int_0^{T_i} \lambda_i(s)ds}. \tag{24}$$

Scheme based on Poisson thinning

We generate points from a Poisson process with intensity $\lambda_0(t)$ (which has a known form)

$$g_1, g_2, g_3, \dots \tag{25}$$

Then we accept / reject the points with probability $\sigma(l(g_k, X))$, we denote the set of rejected points by G and we set $T = g_k$

Putting things into context G represents a set of times where the object hasn't failed.

Density of G and T

Proposition 2

Let $\Lambda_0(t) = \int_0^T \lambda_0(t)dt$, then

$$p(G, T|\lambda_0, I(t)) = \left(\lambda_0(T)\prod_{g\in G}\lambda_0(g)\right) e^{-\Lambda_0(T)} \left(\sigma(I(T))\prod_{g\in G}(1-\sigma(I(g)))\right)$$

A tractable model

Using the previous proposition the model is reformulated as

Tractable model

$$G, T|\lambda_0, I \sim e^{-\Lambda_0(T)}\lambda_0(T)\sigma(I(T)) \prod_{g \in G} \lambda_0(g)(1-\sigma(I(g))) \tag{26}$$

$$I \sim GP(0, K)$$

Some notation

Using the now tractable model, we can perform inference by sampling $G_i | T_i, X_i, \lambda_0, l$ for each (T_i, X_i) and then sample $l | (G_i, T_i, X_i)_{i=1}^n, \lambda_0$.

Before we get into the algorithm, let

$$\mathbf{G} = \cup_{i=1}^n G_i$$
$$\mathbf{T} = \cup_{i=1}^n T_i$$

Algorithm

for q=*1:N* **do**

I
- **for** i=*1:n* **do**
  - $n_i \sim \text{Poisson}(1; \Lambda_0(T_i))$;
  - $\tilde{C}_i \sim U(n_i; 0, \Lambda_0(T_i))$;
  - Set $A_i = \Lambda_0^{-1}(\tilde{A}_i)$;

II
- Set $\boldsymbol{A} = \cup_{i=1}^{n} A_i$
- Sample $l(\boldsymbol{A})|l(\boldsymbol{G} \cup \boldsymbol{T}), \lambda_0$

III
- **for** i=*1:n* **do**
  - $U_i \sim U(n_i; 0, 1)$
  - set $G_{(i)} = \{a \in A_i \text{ such that } U_i < 1 - \sigma(l(a))\}$

IV
- Set $\boldsymbol{G} = \cup_{i=1}^{n} G_i$
- **Update parameters of** $\lambda_0(t)$
- **Update** $l(\boldsymbol{G} \cup \boldsymbol{T})$ and hyperparameter of the kernel.

Let's try breaks this down.

Algorithm I

for *i=1:n* **do**
  $n_i \sim \text{Poisson}(1; \Lambda_0(T_i))$;
  $\tilde{C}_i \sim U(n_i; 0, \Lambda_0(T_i))$;
  Set $A_i = \Lambda_0^{-1}(\tilde{A}_i)$;

The first part uses the **Mapping Theorem** where we construct jump-times.

Algorithm II

6	Set $\boldsymbol{A} = \cup_{i=1}^{n} A_i$
7	Sample $l(\boldsymbol{A}) \vert l(\boldsymbol{G} \cup \boldsymbol{T}), \lambda_0$

Step II involves combining all our sampled jump times into a set

$$\mathbf{A} = \cup_{i=1}^{n} A_i$$

We then sample $l(\mathbf{A})|l(\mathbf{G} \cup \mathbf{T})$ based on a previously fit GP on $\mathbf{G} \cup \mathbf{T}$, i.e. we draw a function from a Gaussian Process, recall at points $\mathbf{A}$ we have a Multivariate-Normal dist.

Algorithm III

for i=*1*:n **do**
  $U_i \sim U(n_i; 0, 1)$
  set $G_{(i)} = \{a \in A_i \text{ such that } U_i < 1 - \sigma(l(a))\}$

Step III is our accept-reject step.

This is where we decide whether to keep points A_i or discard them. Points that are kept are added to a set **G**

Algorithm IV

11 Set $\boldsymbol{G} = \cup_{i=1}^{n} G_i$
12 **Update parameters of** $\lambda_0(t)$
13 **Update** $l(\boldsymbol{G} \cup \boldsymbol{T})$ and hyperparameter of the kernel.

Here we have **G**, so all that's left to do is update the parameters of our hazard rate (using something like MLE), and update our GP

We note our GP can be viewed as a classification problem, where we're trying to classify failure times and non-failure times.

Questions / Concerns?

Censoring

We can accommodate for both left, right and interval censoring.

Right Censoring: continue as usual but don't include censored T_i into **T**

Left / Interval Censoring: We impute a value for T_i and work with that instead

A note on approximations

GPs can be pretty expensive to train; around $\mathcal{O}(n^3)$. So an approximation scheme is proposed to ease some computation.

Experiments

Some experiments were conducted to assess the method.
The models under consideration were

- Cox PHM
- Random Forest Survival
- Anova-DDP
- E-SGP, GP with exponential baseline hazard
- W-SGP, GP with Weibull baseline hazard

Simulations

Simulated $n = 25, 50, 100, 150$ points from

$$p_0(t) = N(3, 0.8^2) \qquad \text{and} \qquad p_1(t) = 0.4N(4, 1) + 0.6N(2, 0.8^2) \quad (27)$$

restricted to $\mathbb{R}^+$.
Data contained the sample points and a covariate indicating which distribution the data was sampled from. Additionally, 3 noisy covariates were added taking values in $[0, 1]$

Simulations

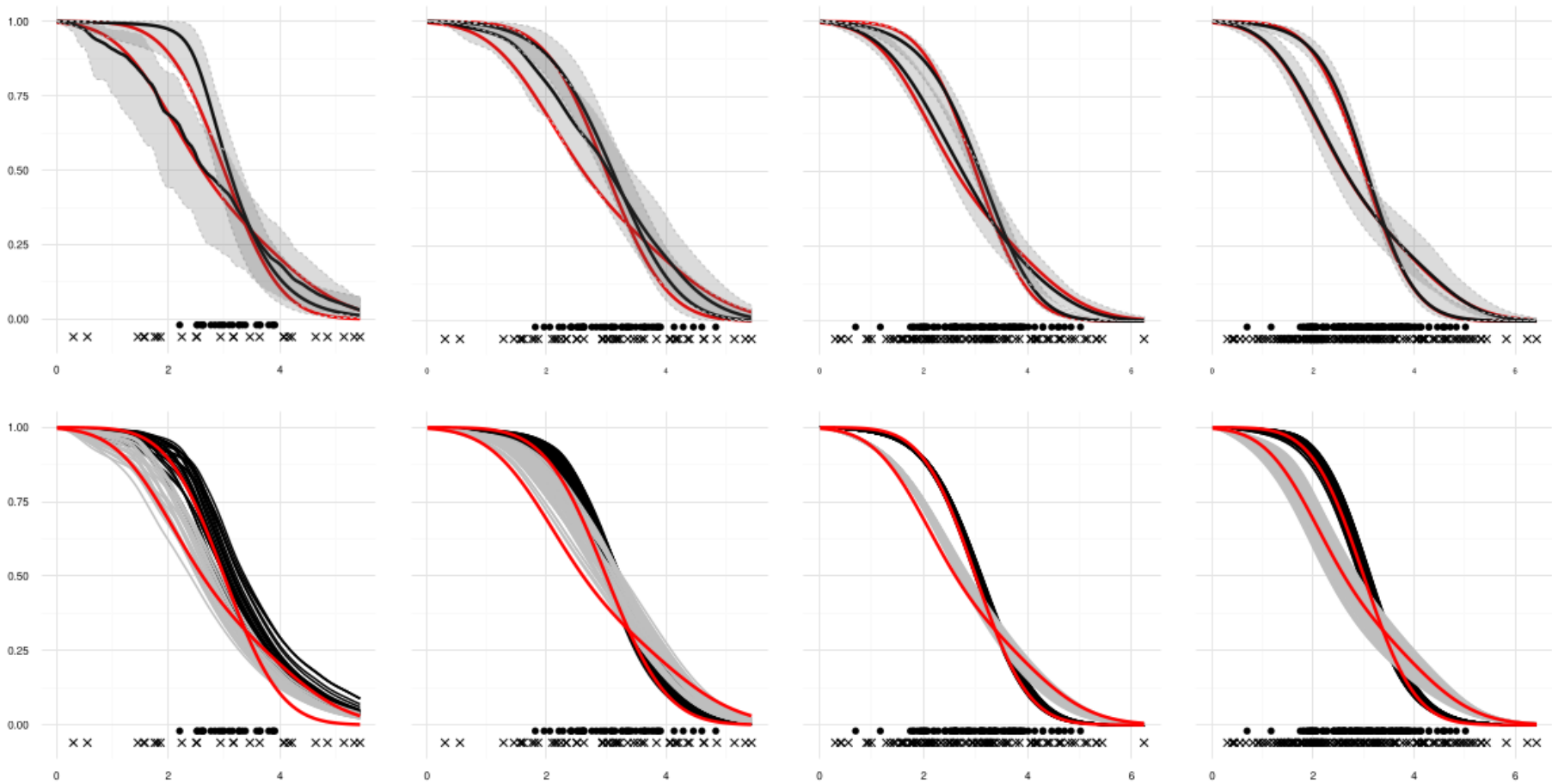


Figure: Weibull-GP, first row noisy data, second row clean. Columns for $n = 25, 50, 100, 150$. In red: true function, black: estimated.

Real-world data

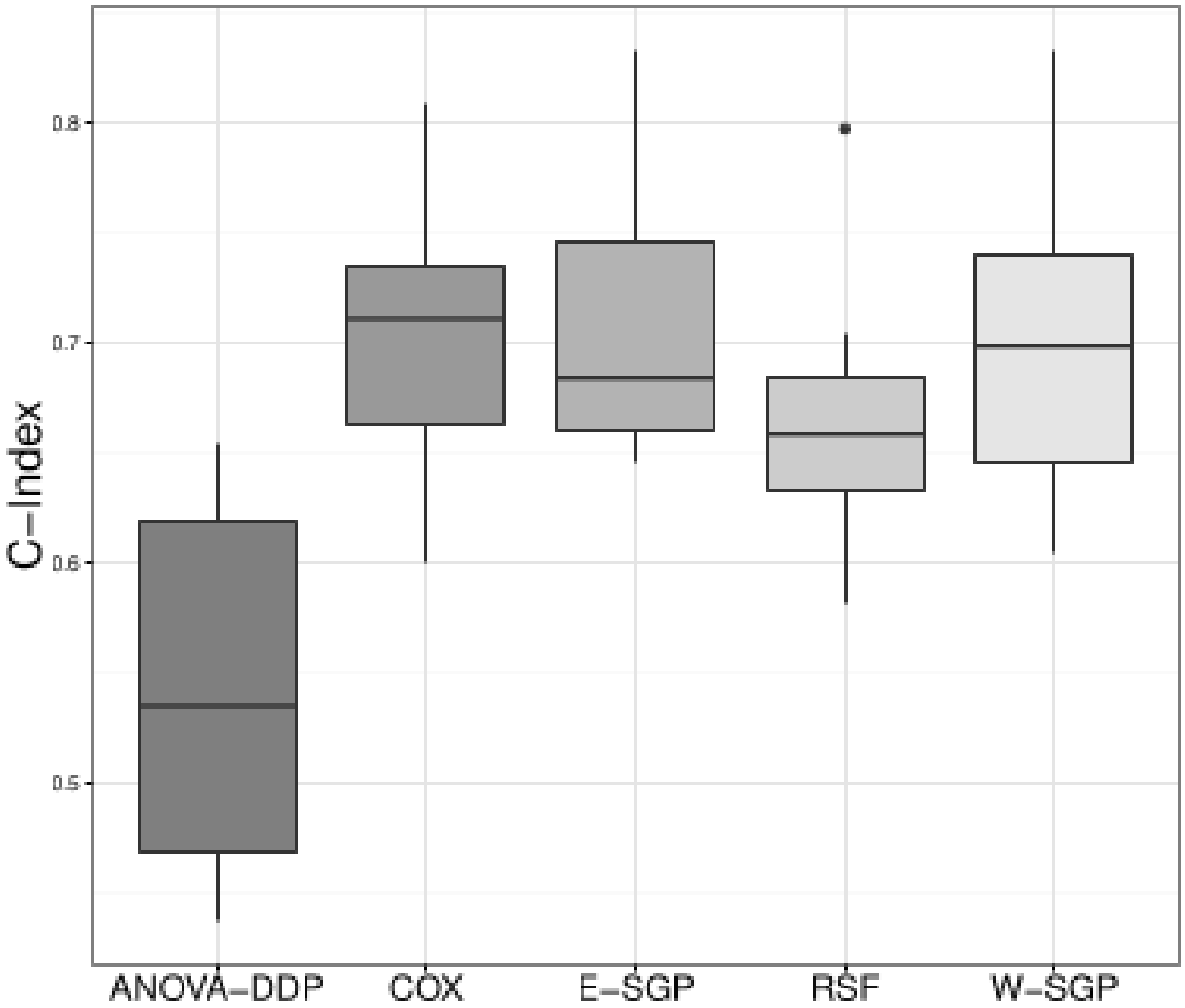
Models were compared using the concordance index. A standard way of comparing survival models.

The `veteran` data from the R package `survival` was used (n=137). Which consists of a randomised trial of two treatment regimes for lung cancer.

There are 5 covariates

- Treatment group
- Age,
- Karnofsky performance score (0-100),
  - 0: dead
  - 50: requires frequent medical care
  - 100: normal/no evidence of disease.
- Indicator of prior treatment
- Months of diagnosis

Real-world data



Thanks for listening

Questions?